EPI 289. Models for Causal Inference – Spring 1 2023
Homework 0 Exercises

Hyewon Jeong
hyewonj@mit.edu

These exercises are due in lecture on Wednesday, January 25th. Use R to complete the following exercises.

1. Read in the **nhefs.xlsx** file from the EPI 289 course website. Show your log to demonstrate that the file was successfully assigned.

```
library(readxl)
df <- read_excel("nhefs.xlsx")
```

```
head(df)
```

seqn	qsmk	death	yrdth	modth	dadth	sbp	dbp	sex	age	···	birthcontrol	pregnancies	cholesterol	hightax82	price71	price82	tax71	tax82
233	0	0	NA	NA	NA	175	96	0	42	···	2	NA	197	0	2.183594	1.739990	1.1022949	0.4619751
235	0	0	NA	NA	NA	123	80	0	36	···	2	NA	301	0	2.346680	1.797363	1.3649902	0.5718994
244	0	0	NA	NA	NA	115	75	1	56	···	0	2	157	0	1.569580	1.513428	0.5512695	0.2309875
245	0	1	85	2	14	148	78	0	68	···	2	NA	174	0	1.506592	1.451904	0.5249023	0.2199707
252	0	0	NA	NA	NA	118	77	0	40	···	2	NA	216	0	2.346680	1.797363	1.3649902	0.5718994
257	0	0	NA	NA	NA	141	83	1	43	···	0	1	212	1	2.209961	2.025879	1.1547852	0.7479248

nhefs.xlsx has been successfully loaded with read_excel function in readxl library. The code segment above shows the head of the loaded dataframe.

2. Sort the data set by the variable **seqn**. Print out the ID number, age, and sex for the first 10 observations.

```
df <- df[order(df$'seqn'),]
df[1:10, c('seqn','age', 'sex')]
```

seqn	age	sex
233	42	0
235	36	0
244	56	1
245	68	0
252	40	0
257	43	1
262	56	1
266	29	1
419	51	0
420	43	0

3. Find the mean **systolic blood pressure** and standard error for men and for women.
Mean systolic blood pressure (SBP) of women is 126.31 (rounded up in second decimal) and 131.25 for men. Standard error was 0.66 for women and 0.67 for men. The details in computation with R code is provided below:

```
df_men <- subset(df, sex==0)
df_women <- subset(df, sex==1)
```

```
mean(df_men$'sbp', na.rm=T)
```

131.246684350133

```
mean(df_women$'sbp', na.rm=T)
```

126.312030075188

```
sd(df_women$'sbp', na.rm=T)/sqrt(length(df_women$'sbp'))
```

0.657332597003731

```
sd(df_men$'sbp', na.rm=T)/sqrt(length(df_men$'sbp'))
```

0.667022407539019

4. What is the mean, 25th percentile, 50th percentile, 75th percentile, and interquartile range of **weight in 1971 (in kilograms)**.

25th percentile, 50th percentile, 75th percentile, IQR are 59.65, 69.4, 79.95, and 20.3, respectively. Calculation is done as below:

```
quantile(df$'wt71', probs = 0.25)
quantile(df$'wt71', probs = 0.50)
quantile(df$'wt71', probs = 0.75)
IQR(df$'wt71')
```

25%: 59.65

50%: 69.4

75%: 79.95

20.3

5a. Using *ifelse* statements, create a new categorical variable corresponding to quartiles of **weight in 1971** as based on the cut-points from Question (4). Give a tabulation of your results.

```
df$'wt71_25' <- ifelse(df$'wt71' < quantile(df$'wt71', probs = 0.25), 1, 0)
df$'wt71_50' <- ifelse(quantile(df$'wt71', probs = 0.25)< df$'wt71' &  df$'wt71'< quantile(df$'wt71', probs = 0.50),
                       1, 0)
df$'wt71_75' <- ifelse(quantile(df$'wt71', probs = 0.50)< df$'wt71' & df$'wt71'< quantile(df$'wt71', probs = 0.75),
                       1, 0)
```

```
table(df$'wt71_25')
table(df$'wt71_50')
table(df$'wt71_75')
```

```
   0    1
1223  406
```

```
   0    1
1229  400
```

```
   0    1
1227  402
```

5b. Create quartiles for **weight in 1971** using *cut* in R. Give a tabulation of your results. Do your results match those of Question (5a)? Why or why not?

The number of individuals assigned for each quantile interval was a bit different from the one that has estimated in 5a. From the tabulated result we can see that the quantile threshold is a bit different from the actual quantile calculated from 4 and that is the reason why we get a bit different result from 5a.

```
df_quart <- cut(df$'wt71', breaks = c(-Inf, quantile(df$'wt71', probs = 0.25), quantile(df$'wt71', probs = 0.50),
                                      quantile(df$'wt71', probs = 0.75), Inf))
```

```
table(df_quart)
```

```
df_quart
(-Inf,59.6] (59.6,69.4]   (69.4,80]   (80, Inf]
        414          402         406         407
```

6. Using *lm* in R, fit a univariate linear regression model for the outcome **weight in 1971** with **number of cigarettes smoked per day in 1971** as the predictor. Report the parameter estimate for cigarettes smoked per day.

```
linear_model <- lm(df$'wt71' ~ df$'smokeintensity', data = df)
```

```
coef(linear_model)
confint(linear_model)
```

(Intercept)	68.734972465389
df$smokeintensity	0.112750159922378

	2.5 %	97.5 %
(Intercept)	67.20484656	70.2650984
df$smokeintensity	0.04818256	0.1773178

```
summary(linear_model)
```

```
Call:
lm(formula = df$wt71 ~ df$smokeintensity, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-35.345 -11.452  -1.718   8.840  99.891

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       68.73497    0.78011  88.109  < 2e-16 ***
df$smokeintensity  0.11275    0.03292   3.425  0.00063 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.68 on 1627 degrees of freedom
Multiple R-squared:  0.007159,  Adjusted R-squared:  0.006549
F-statistic: 11.73 on 1 and 1627 DF,  p-value: 0.0006298
```

7. Create a cross-tabulation between sex and race.

```
table(df$'sex', df$'race', dnn = c("sex","race"))
```

```
   race
sex   0   1
  0 705  94
  1 709 121
```

8. Using *lm* in R, fit a multivariate linear regression model for the outcome **weight in 1971** with **age, sex, and race** as the predictors. From this model, print the observed and predicted values of **weight in 1971** for the first 5 observations. What is the predicted value of weight in 1971 for an individual of age 40, female, and of Black or other race/ethnicity?

```
multivar_linear_model <- lm(wt71 ~ age+as.factor(sex)+as.factor(race), data = df)
```

```
covariate_values <- data.frame('age' = 40, 'sex' = 1, 'race'=1)
predicted_mean <- predict(multivar_linear_model, covariate_values)
```

```
multiv_wt <- predict(multivar_linear_model, df[1:5, c('age','sex', 'race')], type = "response")
multiv_wt
```

1	82.3357584051349
2	76.8713660646623
3	69.499034636176
4	82.1715428935601
5	76.8461021398047

```
predicted_mean
```

1: 69.6000903356066

multivar_linear_model is the linear model with age, sex, race variable. The predicted output of weight in 1971 for the first 5 individals are summarized fisrt, and then prediction result of an individual with given value has printed out.

9. Fit the same model from Question (8) using *glm* in R and compare your results.

```
logistic_model <- glm(wt71 ~ age+as.factor(sex)+as.factor(race), data = df)
```

```
glm_wt <- predict(logistic_model, df[1:5, c('age','sex', 'race')], type = "response")
glm_wt
```

1	82.3357584051349
2	76.8713660646623
3	69.499034636176
4	82.1715428935601
5	76.8461021398047

```
logistic_pred <- predict(logistic_model, covariate_values)
```

```
logistic_pred
```

1: 69.6000903356066

Prediction result was the same for the generalized linear model compared to linear model, which is presented above with R code.

10. Using *glm* with family specified as binomial in R, fit a multivariate logistic regression model for the outcome **asthma diagnosis in 1971** with **age, sex, race, and usual physical activity status (var active)** as the predictors. Print the predicted probabilities of asthma diagnosis for the individuals with the first 5 ID numbers.

```
bin_glm <- glm(asthma ~ age+as.factor(sex)+as.factor(race)+as.factor(active), family=binomial(link = 'logit'),
               data = df)
```

```
asthma_pred <- predict(bin_glm, df[1:5, c('age','sex', 'race', 'active')], type = "response")
asthma_pred
```

1	0.0228775077882475
2	0.0346473369221107
3	0.0426288935696752
4	0.0295294763170105
5	0.0357066704782767

11. (Optional) Create a graph that plots **systolic blood pressure** on the Y-axis and **usual physical activity status (var active)** on the X-axis.

```
plot(df$active, df$sbp, xlab="active", ylab="sbp", main="sbp-sbp")
```